How good are deep models in understanding the generated images?

Ali Borji

Quintic AI, San Francisco, CA

`aliborji@gmail.com`

August 26, 2022

Abstract

My goal in this paper is twofold: to study how well deep models can understand the images generated by DALL-E 2 and Midjourney, and to quantitatively evaluate these generative models. Two sets of generated images are collected for object recognition and visual question answering (VQA) tasks. On object recognition, the best model, out of 10 state-of-the-art object recognition models, achieves about 60% and 80% top-1 and top-5 accuracy, respectively. These numbers are much lower than the best accuracy on the ImageNet dataset (91% and 99%). On VQA, the OFA model scores 77.3% on answering 241 binary questions across 50 images. This model scores 94.7% on the binary VQA-v2 dataset. Humans are able to recognize the generated images and answer questions on them easily. We conclude that a) deep models struggle to understand the generated content, and may do better after fine-tuning, and b) there is a large distribution shift between the generated images and the real photographs. The distribution shift appears to be category-dependent. Data is available at: link.

1 Introduction

Recent deep generative models such as DALL-E 2 [12] and Midjourney[1] have made a big splash. They are capable of synthesizing stunning photo-realistic images for a given input text (*a.k.a.* a prompt), and have inspired many people, in particular the artists. Some researchers have also used these tools to synthesize data for training deep models (*e.g.* [4]). For the most part, the images generated by these systems capture what is included in the input in terms of the objects and their relations. Some studies (*e.g.* [11][2]) have anecdotally and qualitatively inspected these images and have found that they are limited in certain ways. For example, they do not understand the numbers, counting, and negation, have spelling errors, and lack common sense.

On the one hand, deep models such as ResNet [6] are believed to surpass humans in object classification. Here, we test the capability of recent best object classification models on generated images that are easily recognizable by humans. We also investigate the performance of VQA models on answering binary questions on generated images. The outcomes will inform us about the generalization power of deep models.

On the other hand, unlike the significant body of work that has quantitatively evaluated the images generated by GANs [5, 2, 3], little effort has been spent on evaluating DALL-E 2 and Midjourney. The authors of these papers have already used measures such as FID [7] to quantitatively evaluate their systems. However, research has shown that relying on one score is usually not enough to draw strong conclusions. Here, we take a different approach and argue that if generated images are good, then deep models should be able to recognize them. We feed the synthesized images to the best object recognition models and measure the classification accuracy for each object category. We find that, although the average model performance is poor, models score very high and near perfect over some object categories. This indicates that generative models can capture some categories better than others. Visually inspecting the images from the hard categories, reveals that they are indeed hard to recognize by humans (*e.g.* kites). Admittedly, our results should be taken

---

[1] `https://www.midjourney.com/`

[2] See `https://tinyurl.com/yc7u8juf`, `https://tinyurl.com/2p9wku6e`, and `https://www.reddit.com/r/dalle2/`.
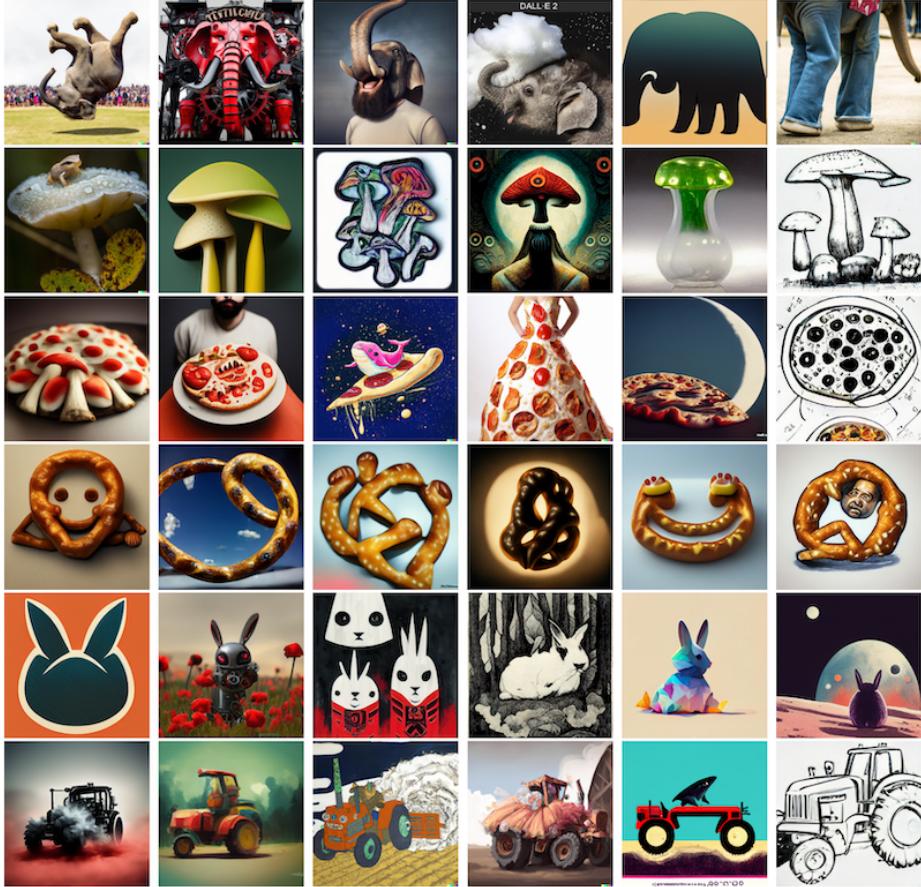
Figure 1: Sample generated images by DALL-E 2 or Midjourney models. Categories in order are: elephant, mushroom, pizza, pretzel, rabbit, and tractor.

with a grain of salt, as classification accuracy may favor a generative model that sacrifices sample diversity in favor of generating high fidelity samples. Generative models are expected to generate a diverse set of high-fidelity images and this is what some scores attempt to measure (*e.g.* FID [7], or Precision-Recall [13]).

We start by curating a dataset of images generated by DALL-E 2 and Midjourney by crawling images from twitter posts as well as Google search. We did not use images that clearly depict faces of people per DALL-E 2 guidelines[3]. Only images that have good quality and are recognizable by humans are selected. We created two sets of images, one for object recognition and another for visual question answering [1]. Sample images from these sets are shown in Fig. 1, and Fig. 4, respectively.

2   Object recognition

We collected 1,862 synthetic images generated by DALL-E 2 and Midjourney across 17 categories (Fig. 1). The number of images per category is shown in the bottom-right panel of Fig. 2. We tested 10 state-of-the-art object recognition models[4], pre-trained on ImageNet, on these images. These models have been published over the past several years and have been immensely successful over the ImageNet benchmark. They include AlexNet [9], MobileNetV2 [14], GoogleNet [15], DenseNet [8], ResNext [19], ResNet101 [6], ResNet152 [6], Inception_V3 [16], Deit [17], and ResNext_WSL [10].

Since some of our classes cover multiple ImageNet classes[5], we had to make some adjustments for computing accuracy. For example, ImageNet has three types of clocks including 'digital clock', 'wall clock', and 'analog clock'. Here, we only have the 'clock' class, containing mostly analog clocks. We

---

[3]https://tinyurl.com/r4xeyhps
[4]Models are available in PyTorch hub: https://pytorch.org/hub/.
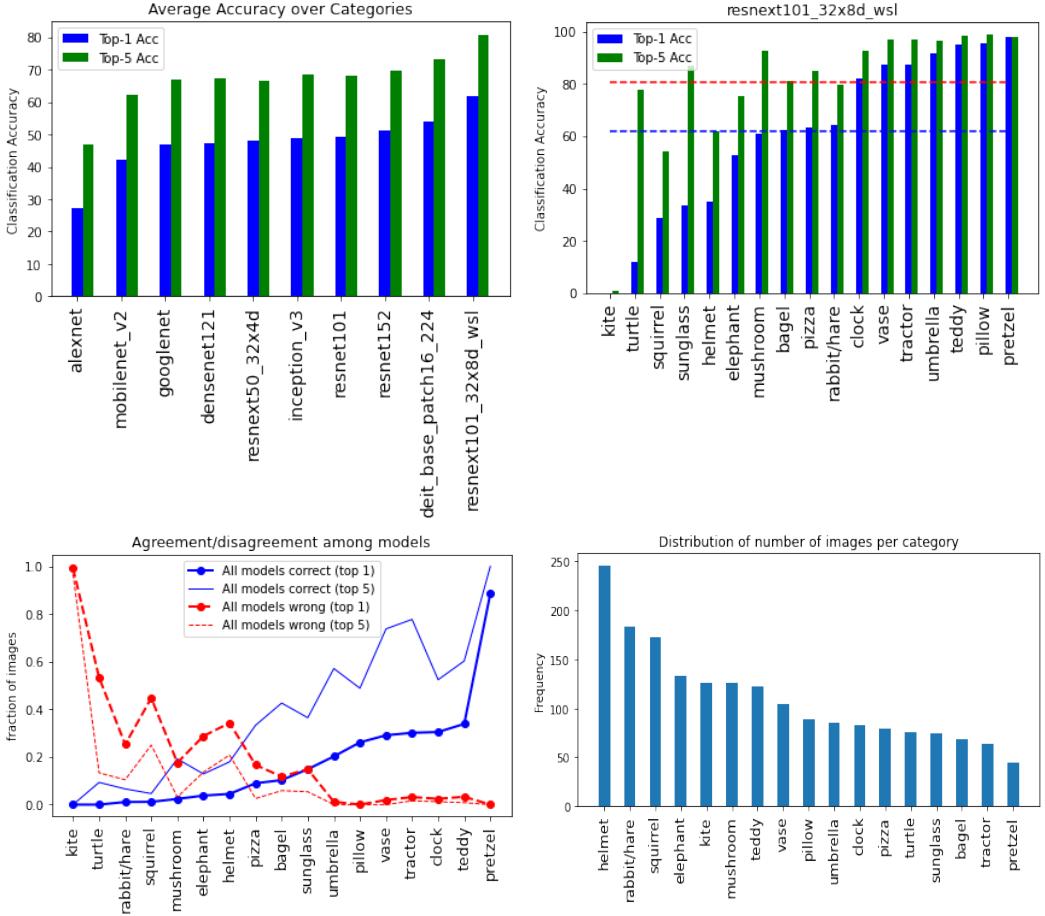[5]https://deeplearning.cms.waikato.ac.nz/user-guide/class-maps/IMAGENET/

Figure 2: Top left: per model performance of the models averaged over 17 categories. Top right: performance of the best model per category. The dashed lines show the average performance. See Appendix A for performance of individual models. Bottom left: fraction of images over which all models fail or they all succeed. Bottom-right: number of images per category.

chose to give the benefit of the doubt to models. A prediction is deemed correct if the ground-truth label is in the set of the words predicted by the model. In the mentioned scenario, if a model predicts 'wall clock', then a hit is counted. If the model predicts 'wall' or anything else, then the prediction would be considered a mistake. The same is true for the top-5 accuracy computation. For example, if the top five model predictions are 'bib', 'necklace', 'toilet seat', 'pick', and 'wall clock', then the prediction is counted as a hit. In practice, first all words in the predicted labels are extracted, and then the prediction is counted as a hit if the ground-truth is in this set. In case of ground-truth having two words (*e.g.* 'toilet seat'), then it should happen in the set of words exactly as it is. Since 'rabbit' and 'hare' classes are very similar, we consider both of them to be true predictions. Notice that this way of accuracy measurement gives an overestimation of the model performance, but it is good enough for our purposes here. Even with this overestimation, as we will show, models still perform poorly.

Results are shown in Fig. 2. Among the models, `resnext101_32x8d_ws` ranks the best, and significantly better than other models. It achieves around 60% top-1 and about 80% top-5 accuracy. This model scores 85.4% top-1 and 97.6% top-5 accuracy over the ImageNet-1k validation set (single-crop). The success of this model can be attributed to the fact that it is trained to predict hashtags on billions of social media images in a weakly supervised manner. The best performance on our data is much lower than the best available performance on the ImageNet validation set which are 91% and 99% corresponding to top-1 and top-5 accuracy[6]. These results suggest that there is a big difference between the distribution of ImageNet images and the distribution of generated images.

---

[6]https://paperswithcode.com/sota/image-classification-on-imagenet

Figure 3: Sample images from the kite category over which the resnext101_32x8d_wsl model fails. An image is considered an error if the ground-truth is not within the top 5 predictions.

According to Fig. 2, the top five most difficult categories for the `resnext101_32x8d_ws` model in order are `kite, turtle, squirrel, sunglass`, and `helmet`. The performance on these categories is below 40%. The kite class is often confused with parachute, balloon, and umbrella classes as shown in Fig. 3. Sample failure cases from the categories along with the predictions are shown in Appendix B. Models often fail on drawings, unusual objects, or images where the object of interest is not unique.

We also computed the fraction of images, per category, over which all models succeed, or they all fail. Results are shown in the bottom left panel of Fig. 2. We noticed that for some categories such as kite and turtle models consistently fail, while for some others such as pretzel and tractor they all do very well. When all models succeed, they are correct at best over 90% of the images (over pretzel category using top-1 acc). These results indicate that models share similar weaknesses and strengths.

3 Visual question answering

Here, we test VQA models on free-form and open-ended visual question answering. We only consider binary questions since in principle, any question can be converted to a binary one on an image. Recent VQA models are able to answer binary questions above 95% accuracy over the VQA-v2 dataset[7], which is astonishing considering the complexity of the questions.

We collected 50 images and formulated a total of 241 questions on them. There are 4.82 questions per image on average. 132 questions have positive answers and 109 have negative answers. Average number of words per question is 5.12 (*i.e.* question length).

To see how well the state-of-the-art VQA models perform on the generated images, we choose the OFA model [18] which is currently the leading scorer on the VQA-v2 test-std set[8]. This model achieves 77.27% accuracy on generated images. To put this result in perspective, this model scores about 94.7% on the VQA-v2 test-std set. There are two reasons why OFA performs lower here a) generated images may contain semantic content that is missing in the training set of the VQA-v2 dataset (*e.g.* 'the astronaut riding the horse'), and b) our questions might be more challenging than

---

[7]https://visualqa.org/
[8]https://paperswithcode.com/sota/visual-question-answering-on-vqa-v2-test-std

Figure 4: Sample images and questions along with the predictions of the OFA model.

VQA-v2 questions. We suspect the first reason is more viable, which is also in alignment with our earlier observations on visual recognition. Sample images and questions along with predictions of the OFA model are shown in Fig. 4. Additional images are shown in Appendix C.

4 Discussion and conclusion

We tested deep models on generated images over two tasks of object recognition and visual question answering. Models perform poorly on these data compared to their performance on real images over which they have been trained on (ImageNet and VQA-v2). We conclude that a) generative models synthesize images for some categories better than other categories, and b) there is a large distribution shift between real images and synthetic images, and this is perhaps why deep models struggle over the latter. We foresee four directions for future research in this area:

1. We did not distinguish between images generated by DALL-E 2 and Midjourney. A quantitative comparison between the two models would be interesting.

2. We tested models on a small test set of generated images. Results over a larger set of images and object classes are likely to provide more insights.

3. It is hard to tell from our results whether low performance on the generated images is due to problems with the images (*e.g.* low fidelity, artifacts, etc), or lack of generalization by the classification models. Visual inspection supports the latter. One way to address this shortcoming is to train and test a deep classifier on generated images. If such a classifier performs well, then it hints towards the high quality of the generated images (and vice-versa).

4. Generative models offer a unique opportunity to automatically generate large scale data for training data-hungry deep models. It would be interesting to see how well models trained on synthetic data generated by DALL-E 2 and Midjourney generalize to real-world data. See [4] as an example in the context of object detection.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] Ali Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019.

[3] Ali Borji. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022.

[4] Yunhao Ge, Jiashu Xu, Brian Nlong Zhao, Laurent Itti, and Vibhav Vineet. Dall-e for detection: Language-driven context image synthesis for object detection. *arXiv preprint arXiv:2206.09592*, 2022.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017.

[8] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[10] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[11] Gary Marcus, Ernest Davis, and Scott Aaronson. A very preliminary analysis of dall-e 2. *arXiv preprint arXiv:2204.13807*, 2022.

[12] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[13] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, pages 5228–5237, 2018.

[14] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[17] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[18] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052, 2022.

[19] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.
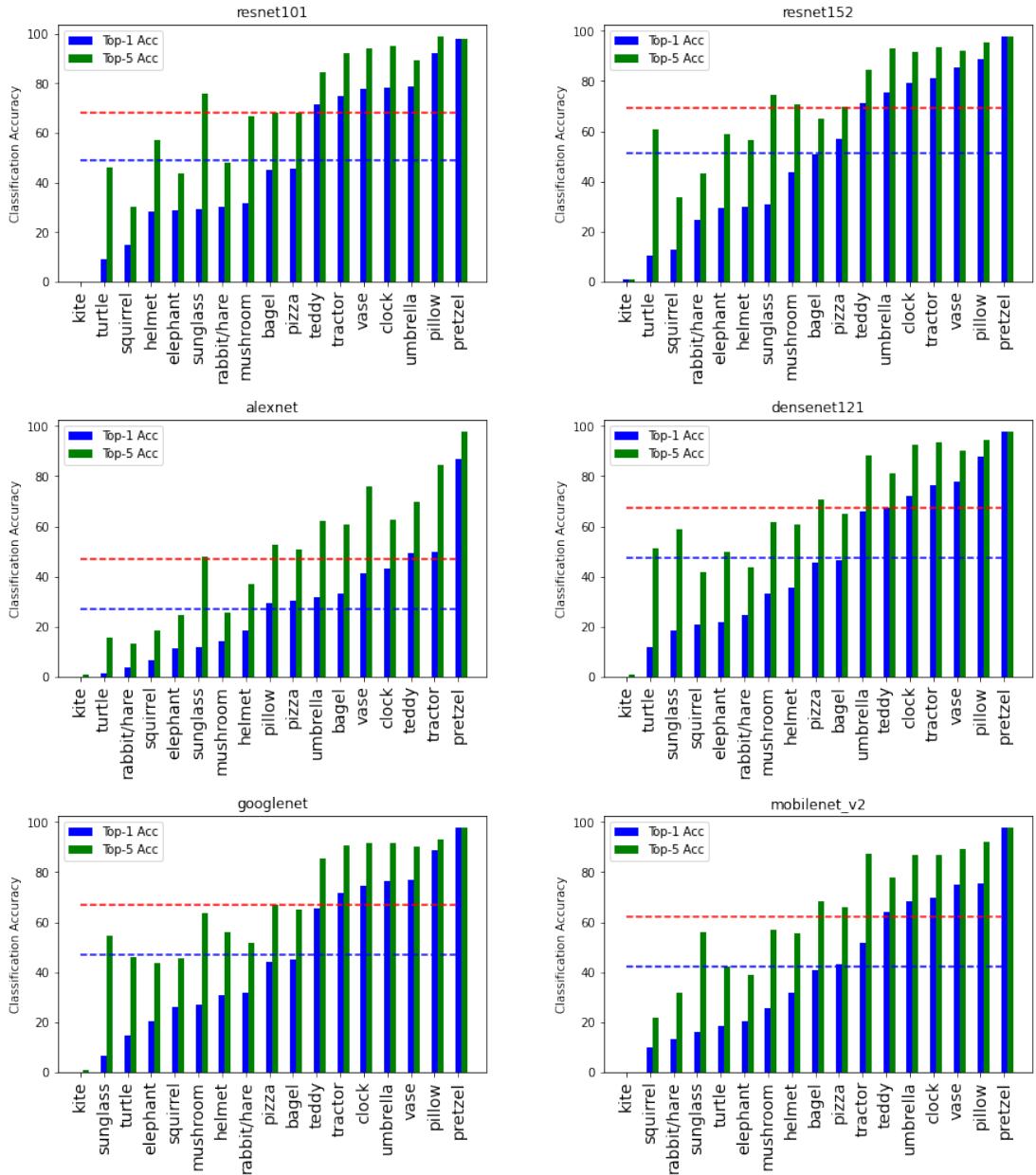
A   Performance of the individual models

Figure 5: Performance of individual models on object detection over generated images.
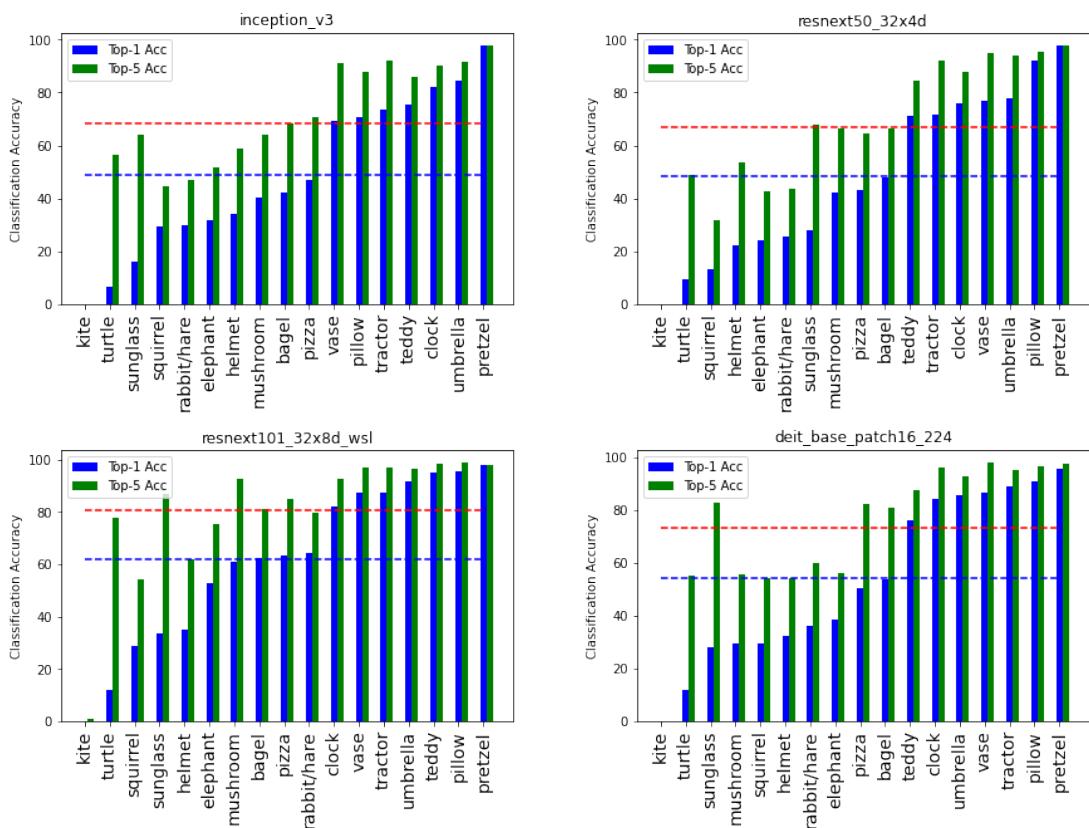
Figure 6: Performance of individual models on object detection over generated images (cnt'd).

B Sample errors made by the resnext101_32x8d_wsl model

Here we show sample errors made by resnext101_32x8d_wsl model over different categories. A image is considered an error if the ground-truth is not within the top 5 predictions.

Figure 7: Predictions of the resnext101_32x8d_wsl model over the bagel category.

Figure 8: Predictions of the resnext101_32x8d_wsl model over the clock category.
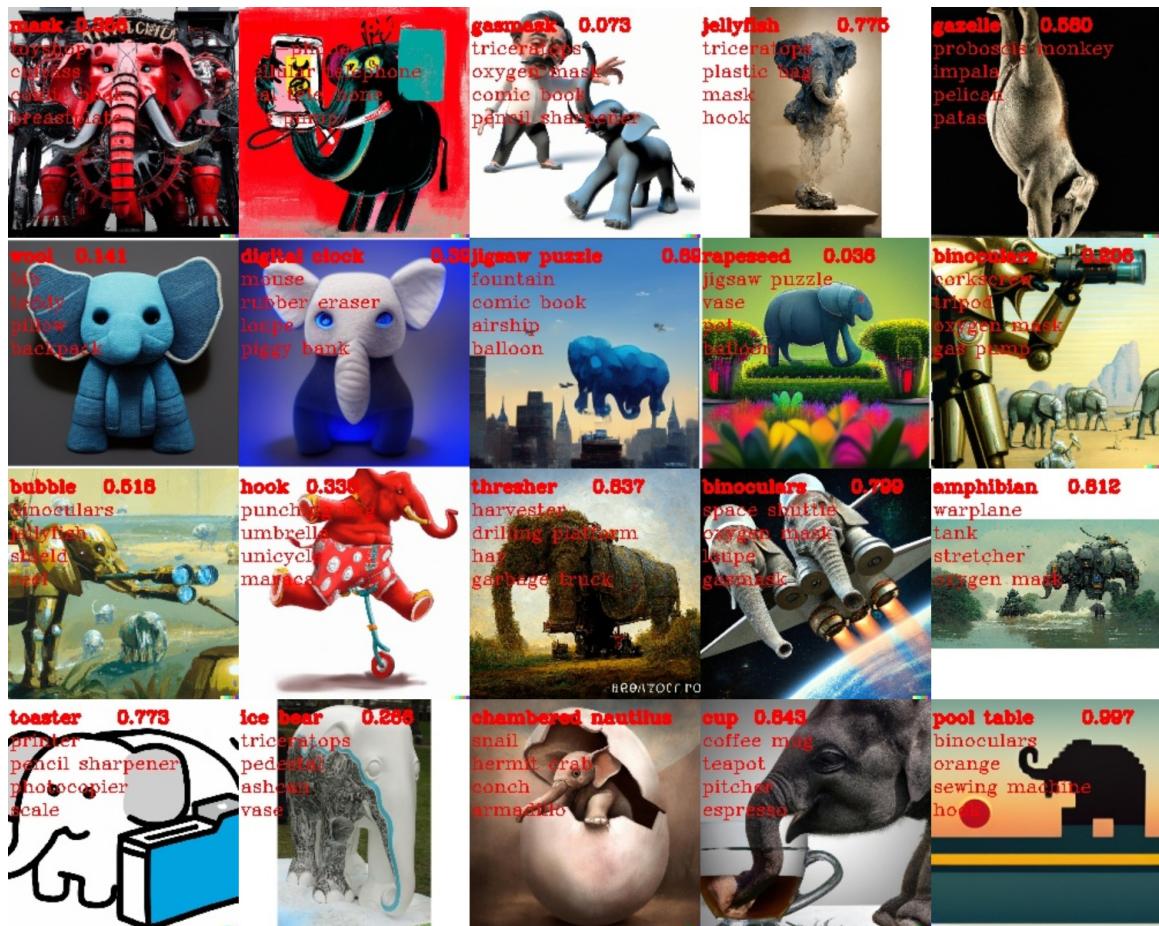
Figure 9: Predictions of the resnext101_32x8d_wsl model over the elephant category.

Figure 10: Predictions of the resnext101_32x8d_wsl model over the helmet category.

Figure 11: Predictions of the resnext101_32x8d_wsl model over the mushroom category.

Figure 12: Predictions of the resnext101_32x8d_wsl model over the pizza category.

Figure 13: Predictions of the resnext101_32x8d_wsl model over the rabbit/hare category.

Figure 14: Predictions of the resnext101_32x8d_wsl model over the squirrel category.

Figure 15: Predictions of the resnext101_32x8d_wsl model over the sunglass category.

Figure 16: Predictions of the resnext101_32x8d_wsl model over the teddy category.

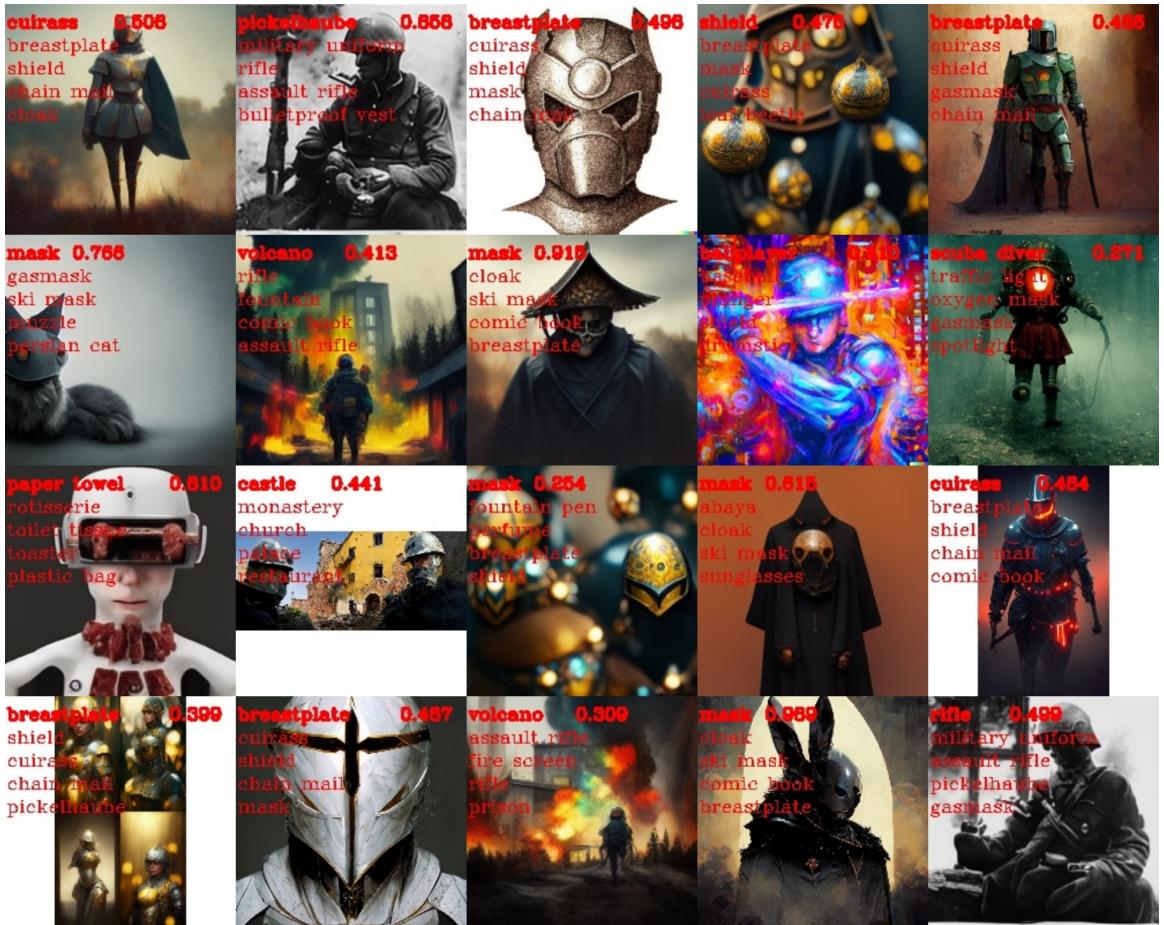Figure 17: Predictions of the resnext101_32x8d_wsl model over the tractor category.
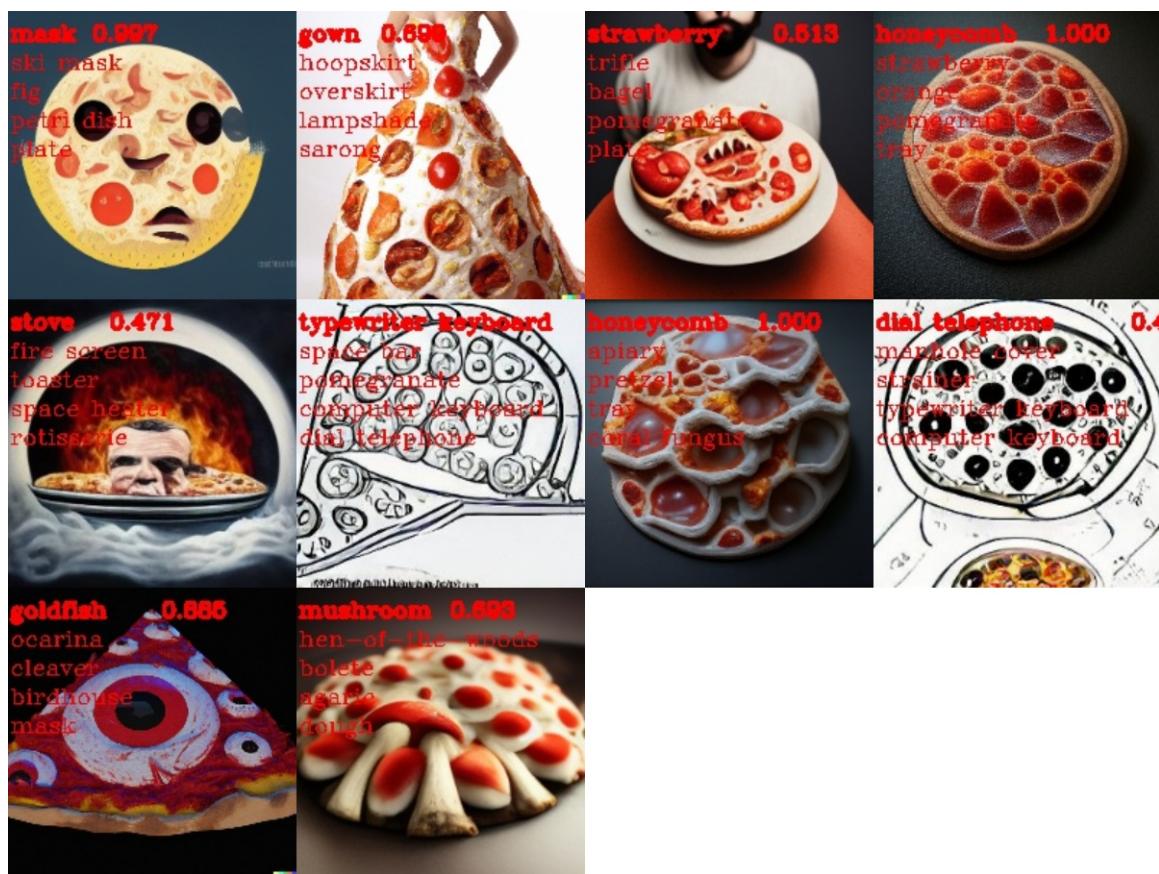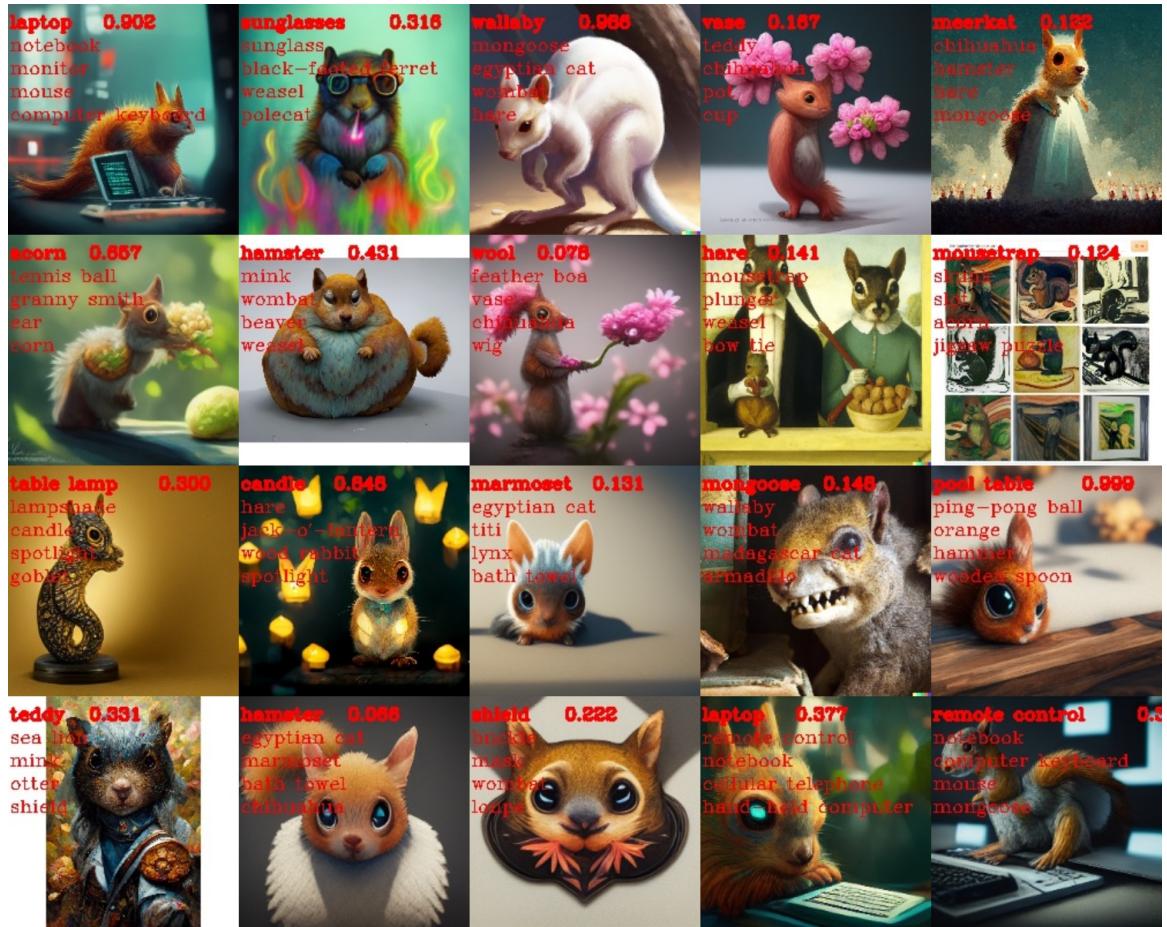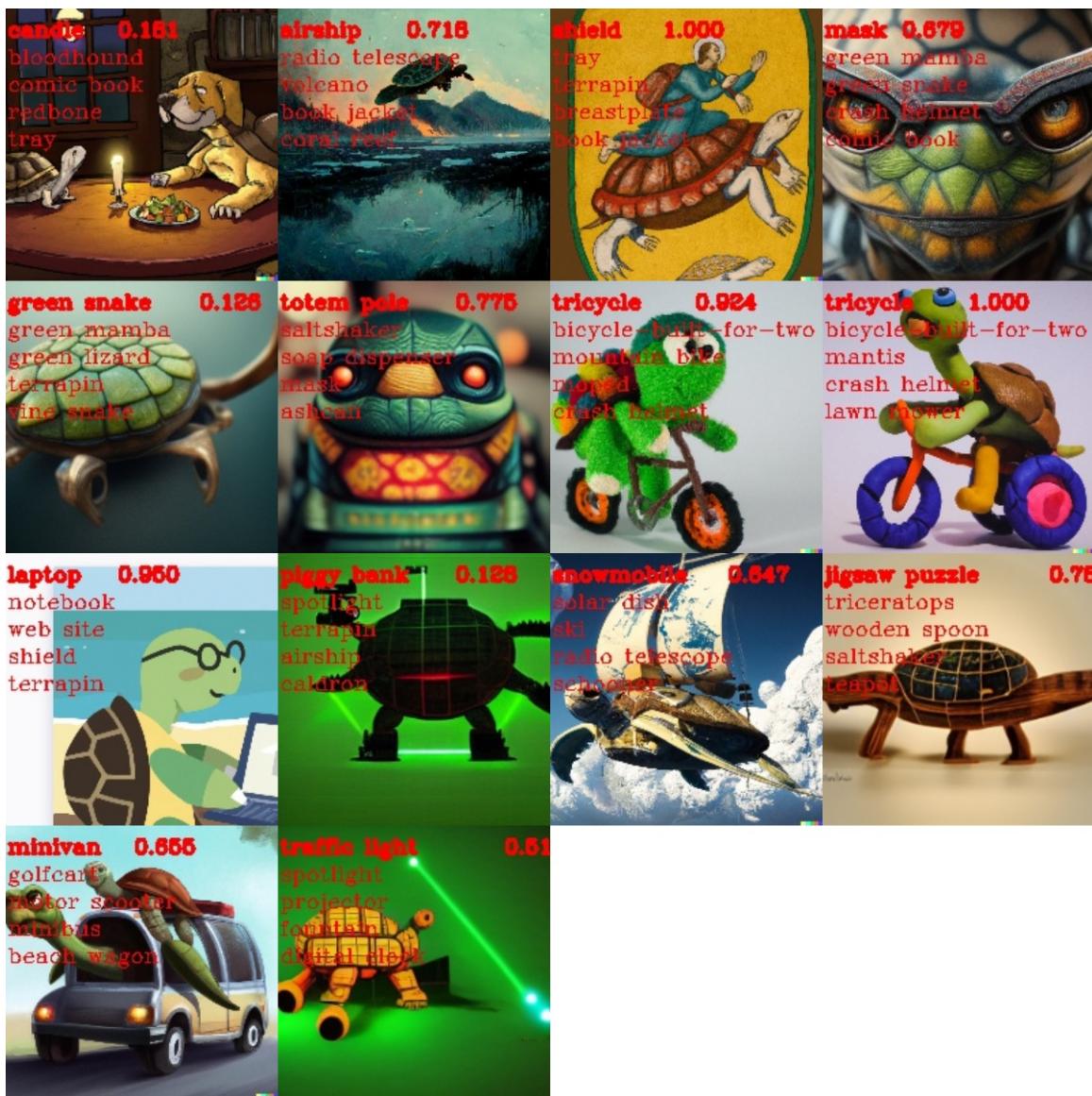

Figure 18: Predictions of the resnext101_32x8d_wsl model over the turtle category.

Figure 19: Predictions of the resnext101_32x8d_wsl model over the umbrella category.

Figure 20: Predictions of the resnext101_32x8d_wsl model over the vase category.
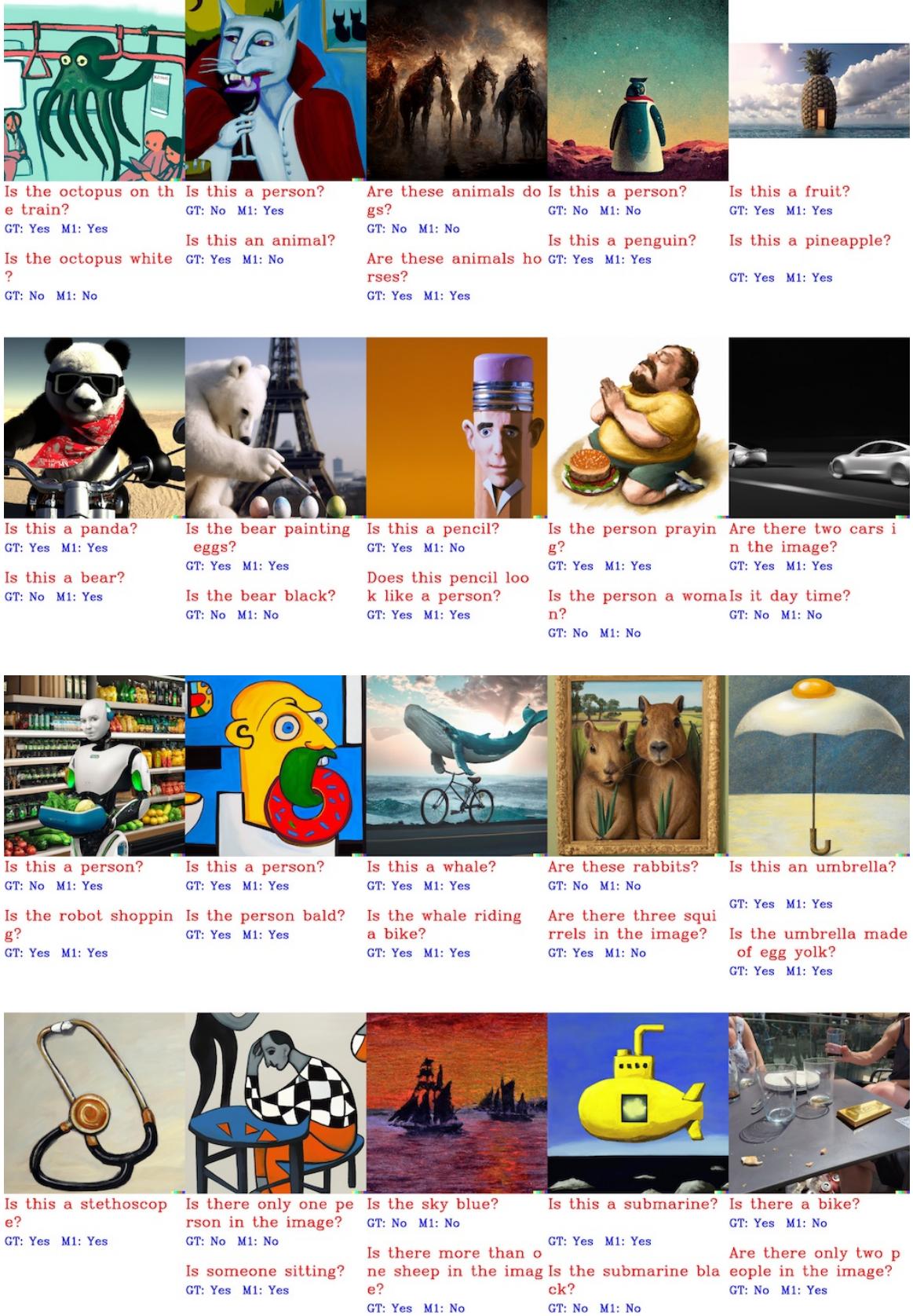
C Additional VQA results

Figure 21: Sample images and questions along with the predictions of the OFA model.

21

Is there a person?
GT: Yes M1: Yes

Is the person on the left side of the image?
GT: No M1: No

Is the rat smoking a cigarette?
GT: Yes M1: Yes

Is this a rabbit?
GT: No M1: No

Is this an elephant?
GT: No M1: No

Does the person in front have a trunk?
GT: Yes M1: No

Is this a dog?
GT: No M1: No

Is this a cat?
GT: Yes M1: No

Is there a bird?
GT: Yes M1: Yes

Is the bird holding a piece of paper with its beak?
GT: Yes M1: Yes

Is this a person?
GT: Yes M1: Yes

Is the full body of the person visible?
GT: No M1: No

Is there a car?
GT: Yes M1: Yes

Is there a dog?
GT: Yes M1: Yes

Is there a crocodile?
GT: Yes M1: No

Is there an octopus?
GT: Yes M1: Yes

Is the panda sitting?
GT: No M1: No

Is the panda walking?
GT: Yes M1: No

Is there a tree?
GT: No M1: No

Is there a person?
GT: No M1: No

Is this a woman?
GT: Yes M1: Yes

Is the woman standing?
GT: No M1: No

Does this look like a skeleton?
GT: Yes M1: Yes

Is the skeleton dancing?
GT: Yes M1: Yes

Are there two people in this scene?
GT: No M1: No

Are there two robots?
GT: Yes M1: Yes

Is this a bird?
GT: Yes M1: Yes

Is this a frog?
GT: No M1: No

Is this a cow?
GT: No M1: No

Is this a camel?
GT: No M1: No

Is there a person?
GT: Yes M1: Yes

Is the person riding a horse?
GT: Yes M1: Yes

Is there a person?
GT: Yes M1: Yes

Is the person a man?
GT: No M1: No

Is this a dog?
GT: No M1: No

Is this a bear?
GT: Yes M1: Yes

Is the person sitting?
GT: No M1: No

Is there a woman?
GT: Yes M1: Yes

Is there a person?
GT: No M1: No

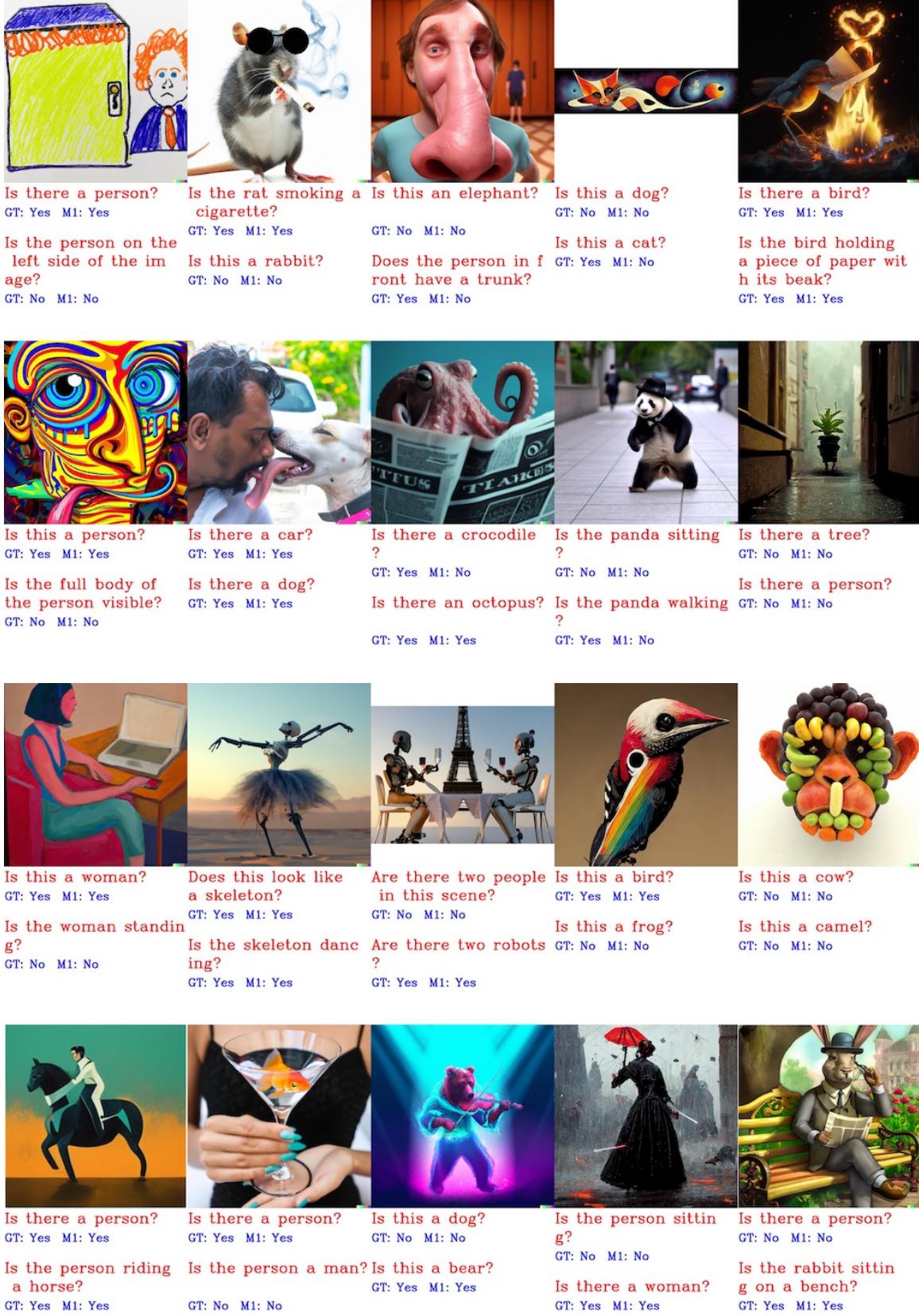Is the rabbit sitting on a bench?
GT: Yes M1: Yes

Figure 22: Sample images and questions along with the predictions of the OFA model (cnt'd).